PRN-21070521035

Name-Kashish Sharma

Section-A2

Summary of attention is all you need

The paper is named Attention is all you need and its authors are Ashish Vaswani , Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

Abstract:

This paper was a major turning point in deep learning research as this paper introduced the transformer architecture which is now used in variety of state of the art models in NLP and many more places.

The model was based on CNN that include an encoder and a decoder.

Introduction:

In sequence modeling such as language modelling and machine translation recurrent neural networks and gated recurrent neural network(a type of Recurrent Neural Network) used to be firmly known as state of the art approaches.Tasks like machine translation, were traditionally handled by sequence to sequence architecture. Sequence-to-Sequence is a neural net that transforms a given sequence of elements, such as the sequence of words in a sentence, into another sequence. Sequence-to-Sequence models consist of an Encoder and a Decoder.

The Encoder takes the input sequence and maps it into a higher dimensional space which is fed into the Decoder which turns it into an output sequence. In this paper authors proposed a new, attention based model (Attention mechanisms are a type of neural network layer that can be added to deep learning models. They allow the model to focus on specific parts of input by assigning different weights to different parts of the input.) and since then attention models have become an essential component in wide range of activities.

This paper introduces transformer, a new deep learning architecture which doesn’t use sequential processing and hence, has a lot of potential for parallel processing.

MODEL:

Encoder:

encoder starts the process by converting words into numerical representation using a embedding layer as computer only understnands numerical data words must be transformed into numbers.transformers does this by using word embedders. word embedder will have similar vectors for words CAT and DOG and different vectors for words like CAT and CAR due to their different contexts. This embedder creats a 512 dimensional vector for each word in the sentence.

Positional encoding:

In transformer models positional encoding plays a crucial role in understanding the order of words in a sentence . while recurrent neural networks orders their words in sentences sequentially transformers does it in parallel or simultaneously. To solve this problem transformer puts positional information into word embedding using positional encoding vector. This vector is combined with the embeddings which helps in transformer to make sure the relative position of words.

Self attention:

The model in this paper uses self attention and multi head attention to understand sentences. self attention focuses on figuring out which words are important for understanding other words in sentences. for example if a sentence contains word IT self attention works on figuring out if IT is referring to an animal or a thing. Multi-head attention runs several self attention processes simultaneously each focusing on different parts of sentences.

Batch and layer normalization:

after the attention block the output vector is added with the orginal vector that was input embedding vector. This combined result is afterwards subjected to a process called normalization which is crucial for model training as it helps reduce training time preve4nts uncontrolled high bias vales etc.

2 common methos for this is-

i)batch normalization

ii)layer normalization

Batch normalization calculates average and variance in a smaller group called batch and then normalizes it to mean closer to 0 and variance closer to 1.

Layer calculation does the same instead it does this for the whole sentence at once ignoring the criteria of batch size.

Feedforward network:

It is shown that in the model of transformer each layer in the encoder as well as the decoder includes a feedforward network. this feedforward network processes the output from one attention layer so it fits better in the following one.

Decoder:

The decoder in the model has different modes in which it operate. During model's training phase the decoder is also in training mode but when the model actually translates a sentence it shifts to testing mode.

for example when translating a sentence from 1 language to another the encoder would be feeded by the material from first language and would produce an output for decoder. Then the decoder would start with start of the sentence token and will make the most likely output until the end of the sentence so basically a part of sentence is feeded to decoder to predict the next word . This method is called as TEACHER FORCING that uses ground truth in place of the output from the prior step.

Masked multi-head attention:

There is a layer in decoder called MASKED MULTI HEAD ATTENTION which is very similar to encoder's attention layer which makes sure the decoder doesn't have access to the future words in the sentence ultimately making sure the model cannot cheat by using the available information.

Encoder-decoder self-attention :

There is a mechanism in transformer model called ENCODER DECODER SELF ATTENTION mechanism which uses encoder's output as KEYS and VALUES while it uses QUERY from decoder. The attention mechanism will calculate attention based on these values.

RESULTS:

Authors in this paper developed 2 different versions of transformers

i) BASE MODEL: it had 65 million parameters that outperformed every existing model in ENGLISH TO GERMAN translation

ii) LARGER MODEL: achieved an excellent BLEU score of 41.8 for ENGLISH TO FRENCH translation